

# ***Cyber Dumpster Diving – creating new software systems for less***

**Ian Gorton, Lab Fellow,  
Group Lead, Data Intensive Scientific Computing,  
Computational Sciences and Math Division  
Pacific Northwest National Lab**



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# PNNL

- ▶ Department of Energy Science Lab
  - Physical and fundamental sciences
  - National security
- ▶ 4500+ people
- ▶ Business volume of over \$1b per annum
- ▶ Large scale experimental facilities, e.g.
  - Environmental Molecular Sciences Lab (EMSL)
  - 161 Tflop supercomputer

**The most radical possible solution  
for constructing software is not to  
construct it at all.**

*Fred Brooks: No Silver Bullet: Essence and Accidents of Software Engineering*



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# DISC@PNNL

## ► Data Intensive Scientific Computing

- User platforms
- Data management
- Tool integration
- Workflows
- Provenance

## ► Applications in e.g.

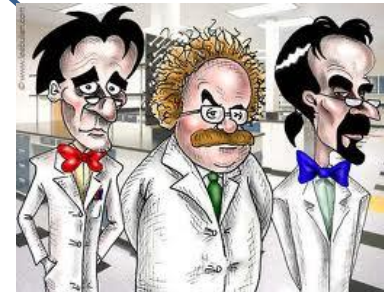
- Bioinformatics
- Climate modeling
- Carbon sequestration
- Subsurface modeling



## High Performance Computing



## Scientific User Platforms



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

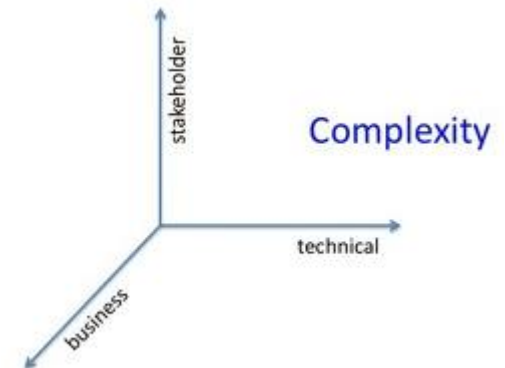
# The middle is a hard place ...

## ► Requirements

- Need to understand science domain
- Need to understand HPC
- Difficult to define, constant refinement, negotiate, negotiate
- “The hardest single part of building a software system is deciding precisely what to build.”

## ► Design

- Conflicting quality requirements
- Complex, heterogeneous technologies
- Large data
- Proliferation of tools, variable quality



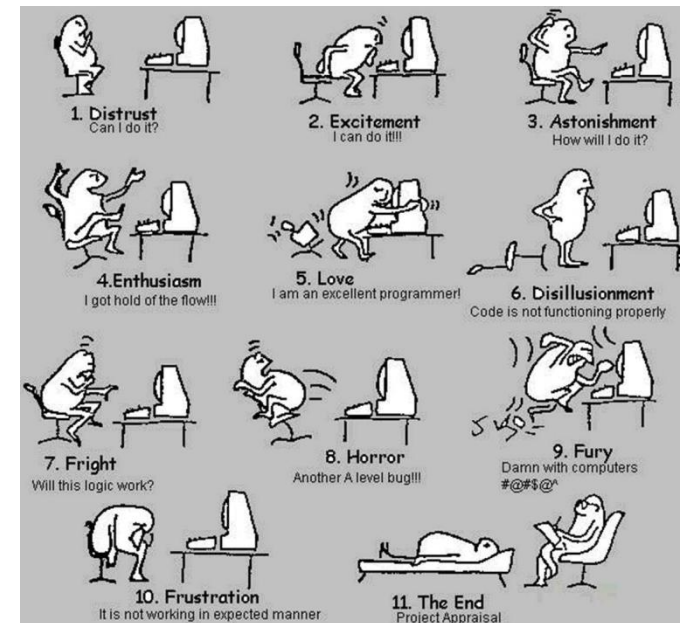
**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*



# Project Funding Profiles

- ▶ Typically fixed amounts
  - What can we build with X dollars?
  - Fixed amounts per year, 1-3 year lifecycle
- ▶ Limited funding
  - From .25 to 10 team size per year
  - 1-2 people per year most common
- ▶ High expectations
  - Scientists think 'software is easy'
  - it's just coding, right?



Pacific Northwest  
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

# Some Examples

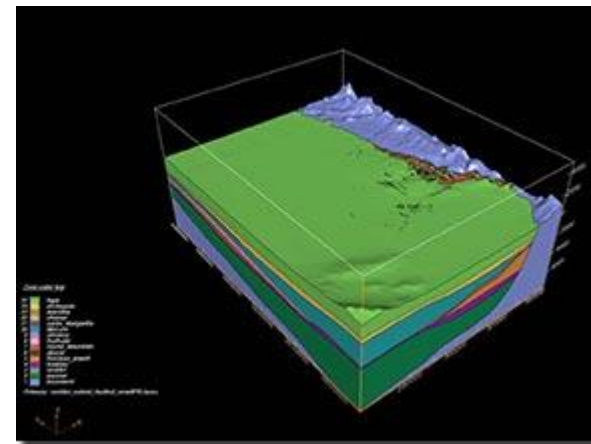
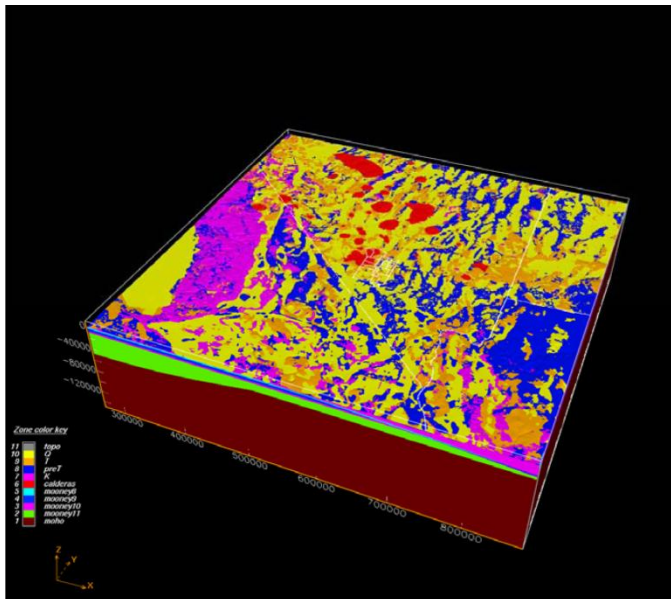
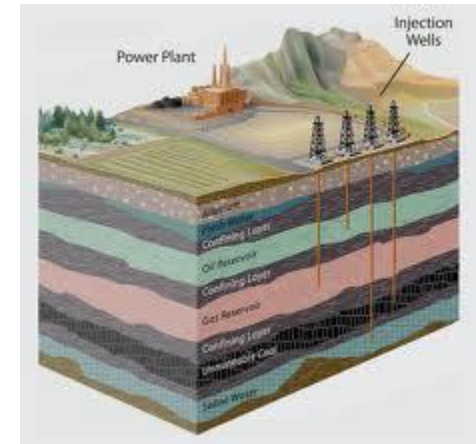
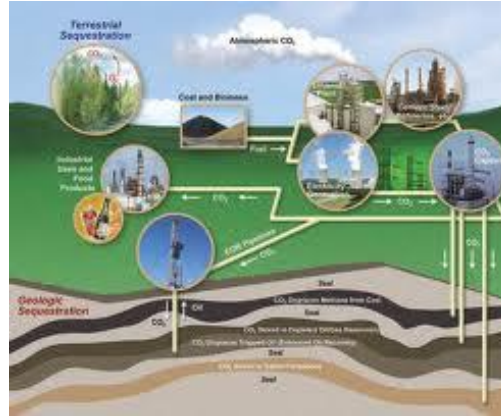


**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*



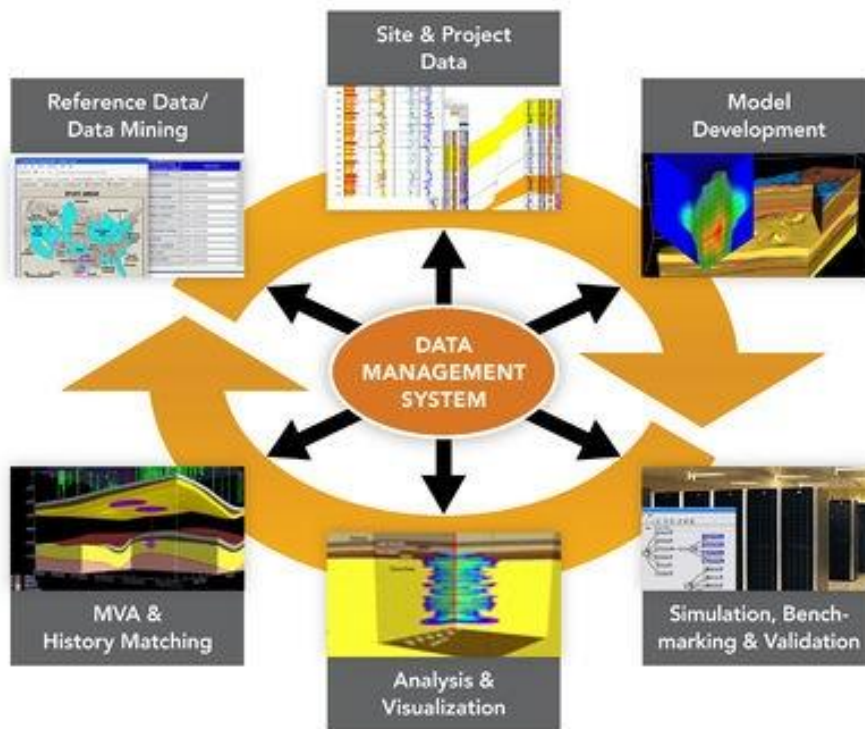
# Carbon Sequestration (Storage)



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Geological Sequestration Software Suite (GS3)



- ▶ Large-scale, complex data
  - Experimental
  - HPC Simulation inputs/outputs
  - Multiple realizations for UQ
- ▶ Long-lived projects
  - Modeling
  - Analysis
  - Monitoring (100+ years)

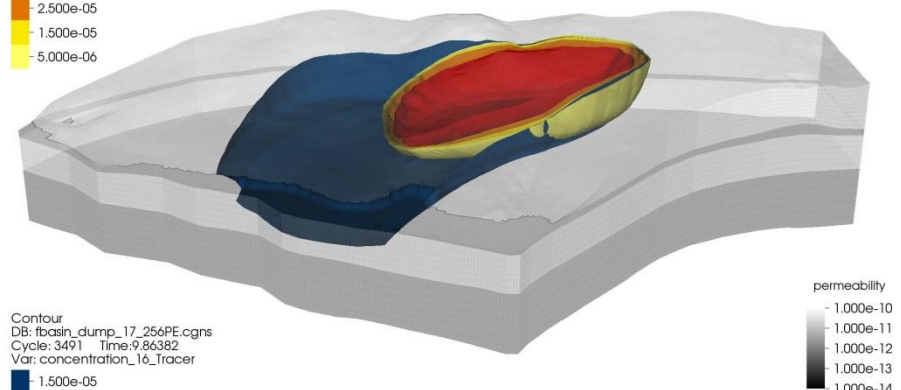
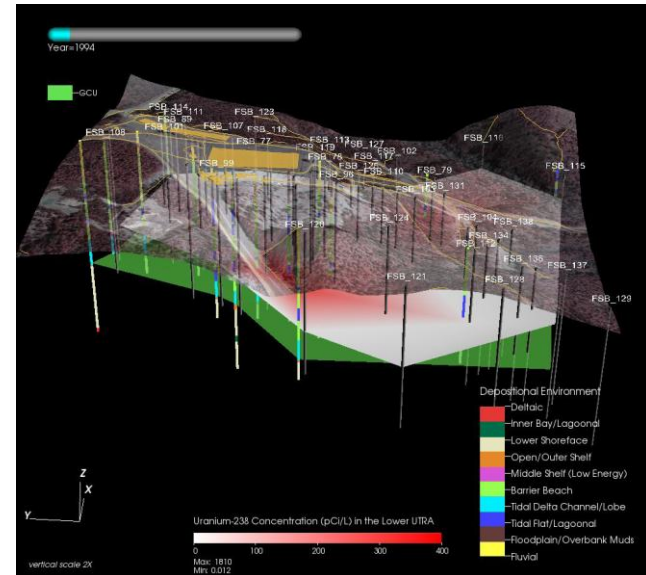


**Pacific Northwest**  
NATIONAL LABORATORY

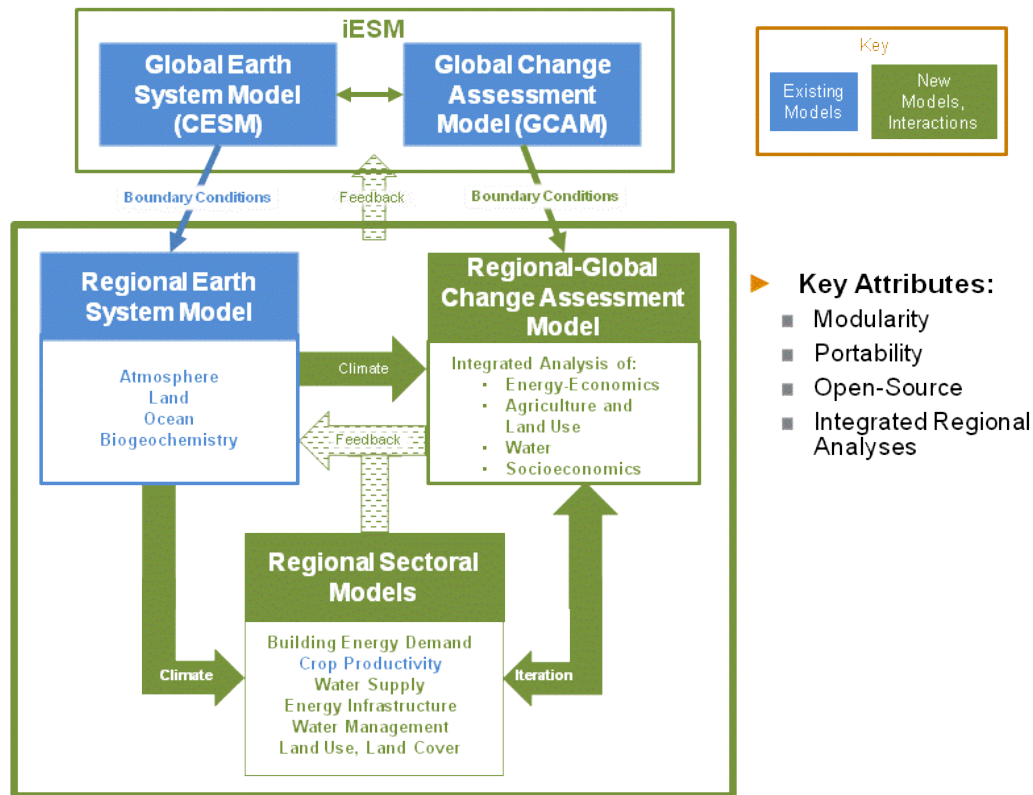
*Proudly Operated by Battelle Since 1965*

# Advanced Simulation Capability for Environmental Management (ASCEM)

- ▶ A State-of-the-art tool for predicting contaminant fate and transport through natural and engineered systems
  - Open source
  - Modular and extensible
  - 'born' parallel for execution on emerging architectures.
- ▶ A User Platform to manage data, create models and analyze results



# Integrated Regional Earth System Model



5





***A powerful, usually legal, source of information that isn't seriously defended because of social taboos.***



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*



# ‘Write-as-little-as-possible’ Reuse

## ► Approach:

- Leverage open source frameworks and tools
- Extend to support science applications
- Generalize to support multiple science domains

## ► Requires:

- Careful technology selection
- Creative design
- Robust architectures



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# An simple example - bioinformatics



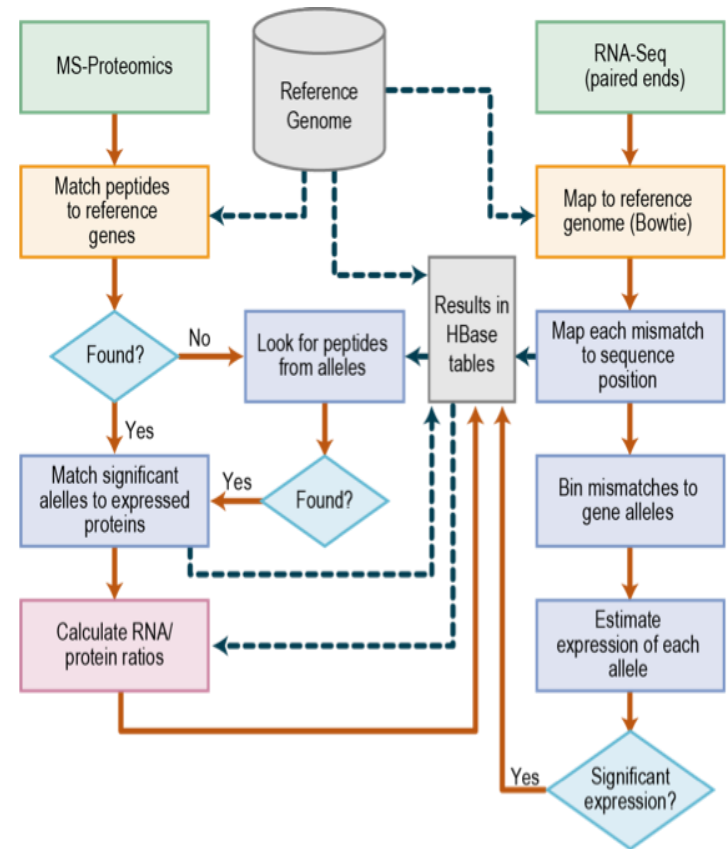
**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*



# Bioinformatics Workflows

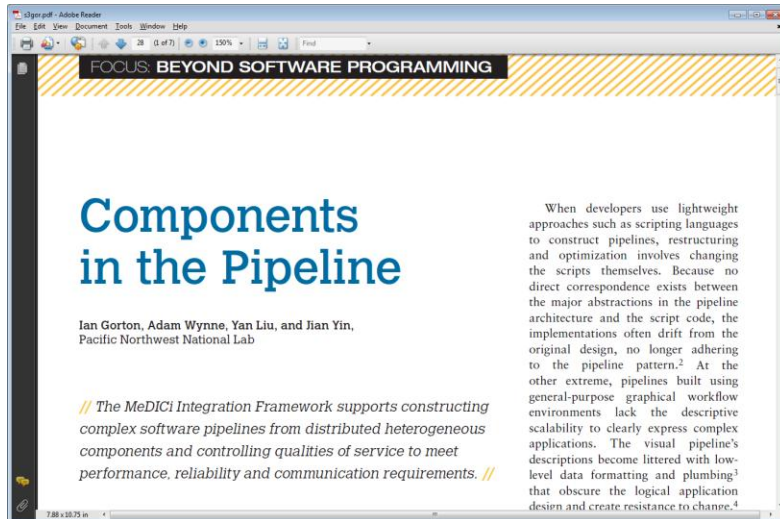
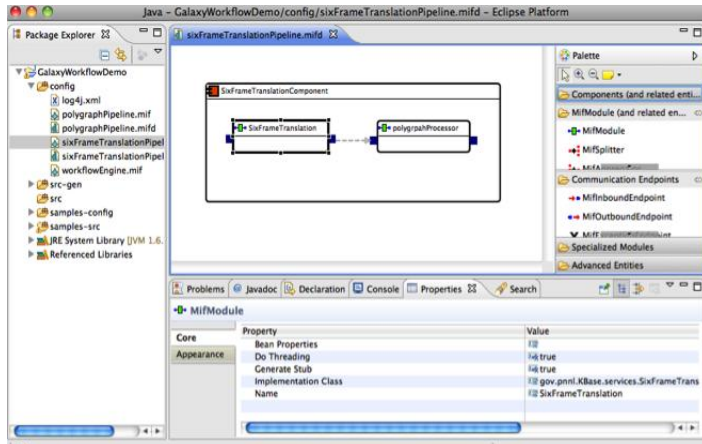
- ▶ Biologists want to create computational workflows:
  - Integrate disparate tools and data
  - Span multiple execution platforms
  - Execute reliably and efficiently
  - Capture provenance
- ▶ Workflow tools for biology
  - Galaxy
  - Taverna
- ▶ Limitations in handling large scale data and computations



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Computational Pipeline Infrastructure



<http://www.computer.org/portal/web/csdl/doi/10.1109/MS.2011.23>

# **Velo – Knowledge Management for Modeling and Simulation**

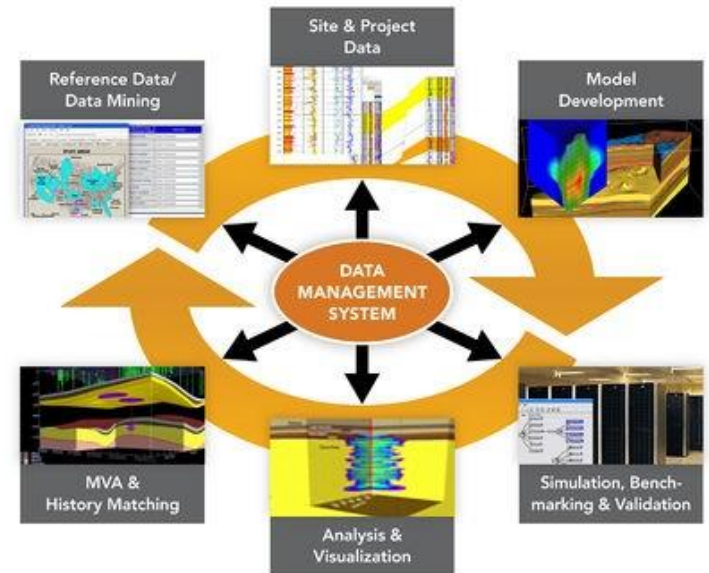


**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Supporting Carbon Sequestration Modeling

- ▶ Requirements
  - Collaboration
  - Sharing data
  - Metadata management
  - User-driven customization
  - Extensibility
  - Model versioning
  - Provenance
  - Robust, scalable
- ▶ Small project, team ~1.75 people, 3 years



# Cyber Dumpster Diving

- ▶ Open source
- ▶ Candidate technology assessments:
  - Quality of docs
  - Release schedule
  - Community scope
  - APIs
  - Code/architecture
  - Install and workout, simple tests



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Feature-Reuse Matrix

Feature	Solution	Notes	Reuse
Collaboration	Mediawiki	Core wiki features support this	100%
Sharing data	Mediawiki Alfresco	Requires integration of MW and Alfresco	60%
Metadata management	Mediawiki Alfresco	Requires customization of MW and Alfresco basic features	80%
User-driven customization	Mediawiki	Core wiki features support this	100%
Extensibility	Mediawiki Alfresco	APIs support extension, but requires design of exact integration mechanisms	20%
Model versioning	Mediawiki Alfresco	Minor extensions for MW/Alfresco capabilities	75%
Provenance	Mediawiki	Some for free in MW, but advanced features need developing	20%
Role-based Security	Halo ACL	Mediawiki extension	100%

# GS3 Examples - Semantic Capabilities - Metadata Extraction

## ► Metadata:

- Generic information e.g. file size, owner, preview/thumbnails
- Specific to the file type, e.g. keywords, geographic location

## ► Metadata is searchable

## ► Extensible architecture for custom data types ingest pipelines, e.g.

- Simulation outputs
- Spreadsheets
- Input files

The screenshot displays the 'Velo Demonstration' web application. The top navigation bar includes 'Home', 'Browse', 'Tools', 'Misc. Tools', and 'Account Links'. The main content area shows a file browser view for the path '/refdata/illinois Basin/Zhou et al 2009.pdf'. A PDF icon is visible, and a 'PDF First Page Preview' section shows the title 'Modeling Basin- and Plume-Scale Processes of CO2 Storage for Full-Scale Deployment' and an abstract. To the right, a table titled 'Top Category/Keyword Matches' lists various geological and physical properties with their respective counts.

Top Category/Keyword Matches	
Rock Name (32)	Sandstone (17) Granite (8) Shale (7)
Geologic Formation Name (156)	Mt. Simon (119) Eau Claire (34) St. Peter (3)
Rock Property (188)	Permeability (93) Salinity (29) Porosity (22) Compressibility (19) Depths (13) Thick (12)
Data Identifiers (75)	Core (61) Wells (9) Geophysical (3) Seismic (2)
Geologic Sequestration Unit (29)	Caprock (25) Reservoir (4)



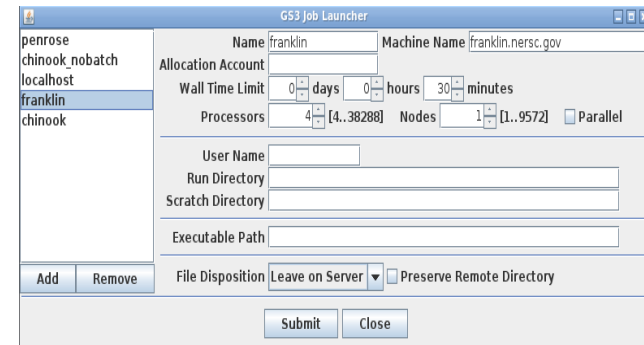
**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

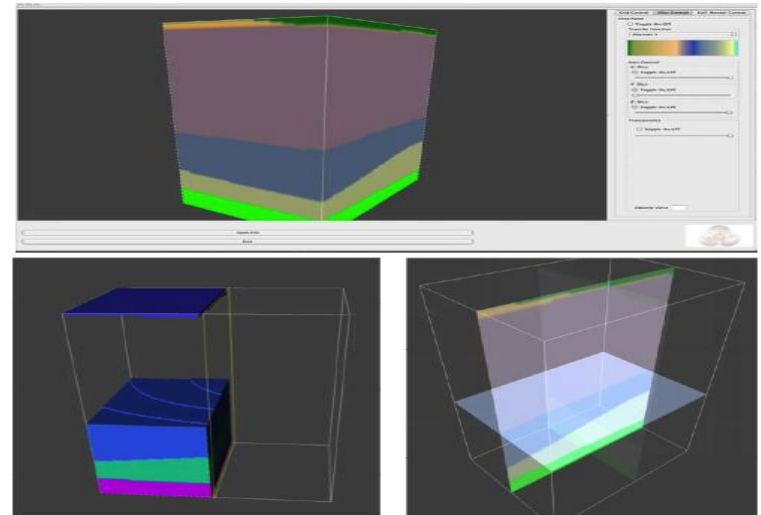


# GS3 Examples - Tool Integration

- ▶ Mediawiki plugins
- ▶ 'Black box' tools
- ▶ External 3<sup>rd</sup> party tools



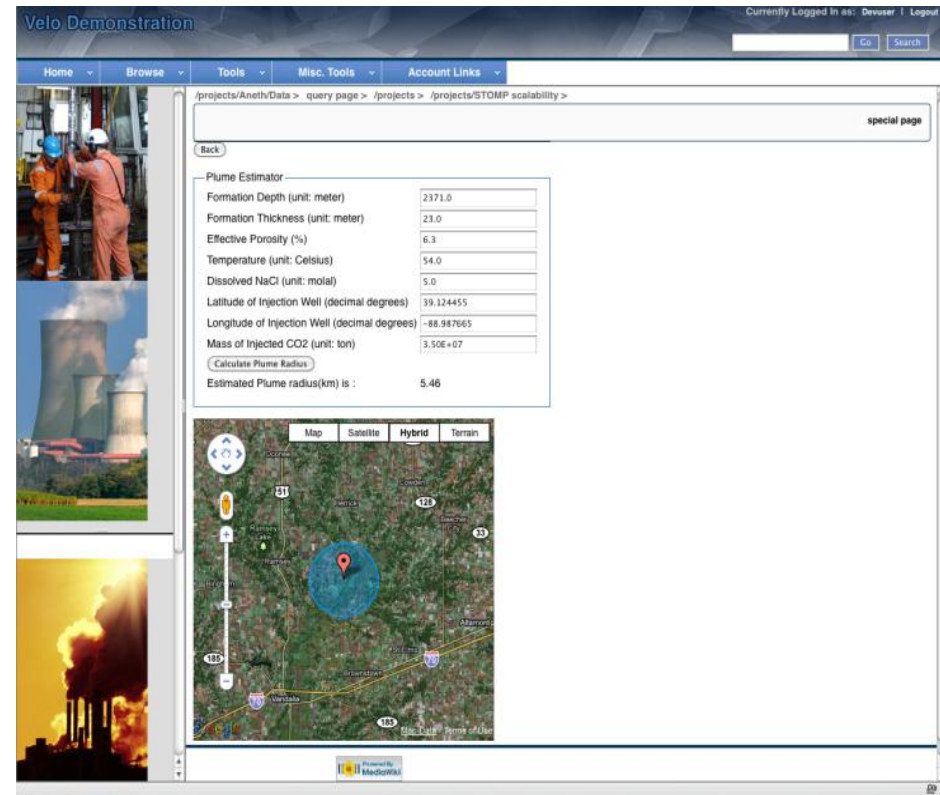
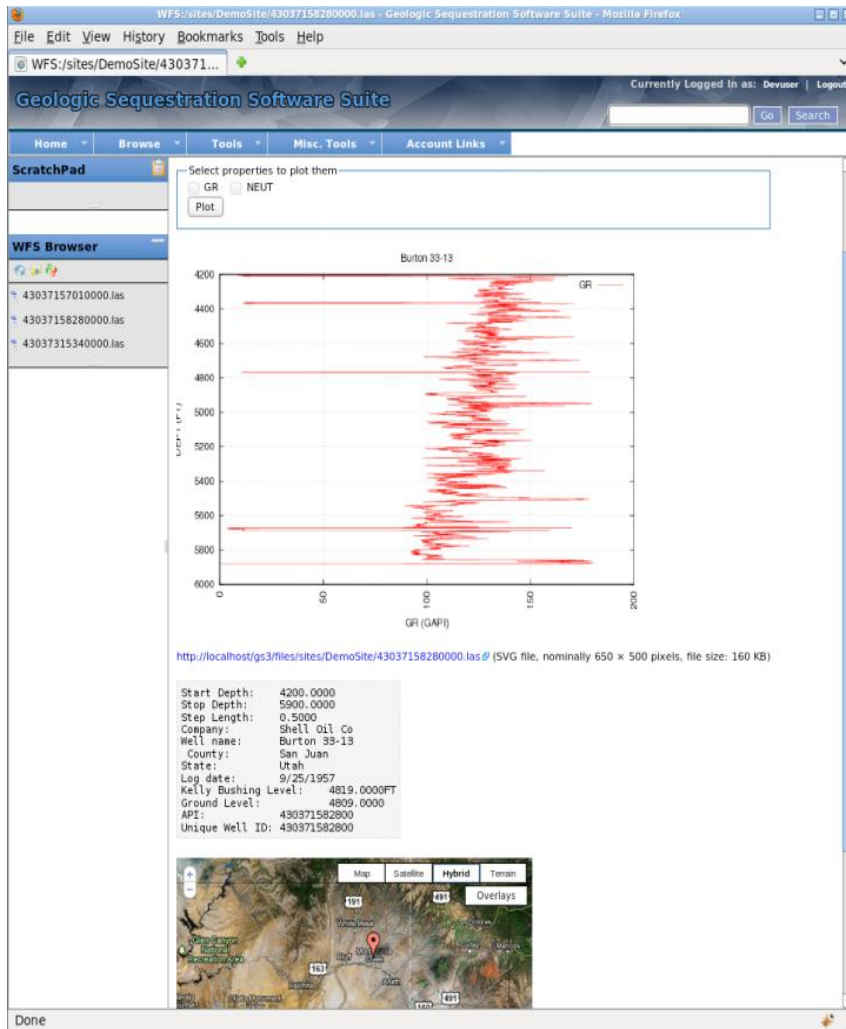
The image shows a screenshot of the 'GS3 Job Launcher' window. It contains several input fields and controls for configuring a job. On the left, there is a list of machine names: 'penrose', 'chinook\_nobatch', 'localhost', 'franklin' (which is highlighted), and 'chinook'. The main area has fields for 'Name' (set to 'franklin'), 'Machine Name' (set to 'franklin.nersc.gov'), 'Allocation Account' (empty), 'Wall Time Limit' (0 days, 0 hours, 30 minutes), 'Processors' (4, with a range of [4..38288]), and 'Nodes' (1, with a range of [1..9572]). There is a 'Parallel' checkbox. Below these are fields for 'User Name', 'Run Directory', 'Scratch Directory', and 'Executable Path'. At the bottom, there are 'Add' and 'Remove' buttons, a 'File Disposition' dropdown menu (set to 'Leave on Server'), a 'Preserve Remote Directory' checkbox, and 'Submit' and 'Close' buttons.



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# GS3 Examples – Tool Plugins



# GS3 Examples – Black box Tool Plugins

GS3 Model Attributes: General Site Description

Location Geography Reservoir Category **Target Formations** Seal Formations

Number of Target Formations within Reservoir: 1

**Target Formation #1**

Target Formation Name: Mt Simon Geologic Age: Cambrian

Target Formation Rock Types (check all that apply):

☒ Sandstone ☐ Limestone ☐ Dolomite ☐ Shale ☐ Coal Seam ☐ Basalt

Other (Specify):

Depositional Environment (check all that apply):

Continental: ☐ Alluvial ☐ Aeolian ☐ Fluvial ☐ Lacustrine

Transitional: ☐ Deltaic ☐ Tidal ☐ Lagoonal ☐ Beach

Marine: ☒ Shallow Water ☐ Deep Water ☐ Reef

Others: ☐ Evaporite ☐ Glacial

Sequestration Trapping Mechanisms (check all that apply):

☒ Dissolution and Diffusion ☐ Physical Containment ☐ Mineralization ☐ Residual Saturation

Other (Specify):

Target Reservoir Depth and Thickness:

Top Depth: Min: Max: Mean: 6705 ft

Bottom Depth: Min: Max: Mean: 9241 ft

Thickness: Min: Max: Mean: 2536 ft

Estimated Fracture Gradient: 0.8 psi/ft

Estimated Fracture Opening Pressure: 5200 psi

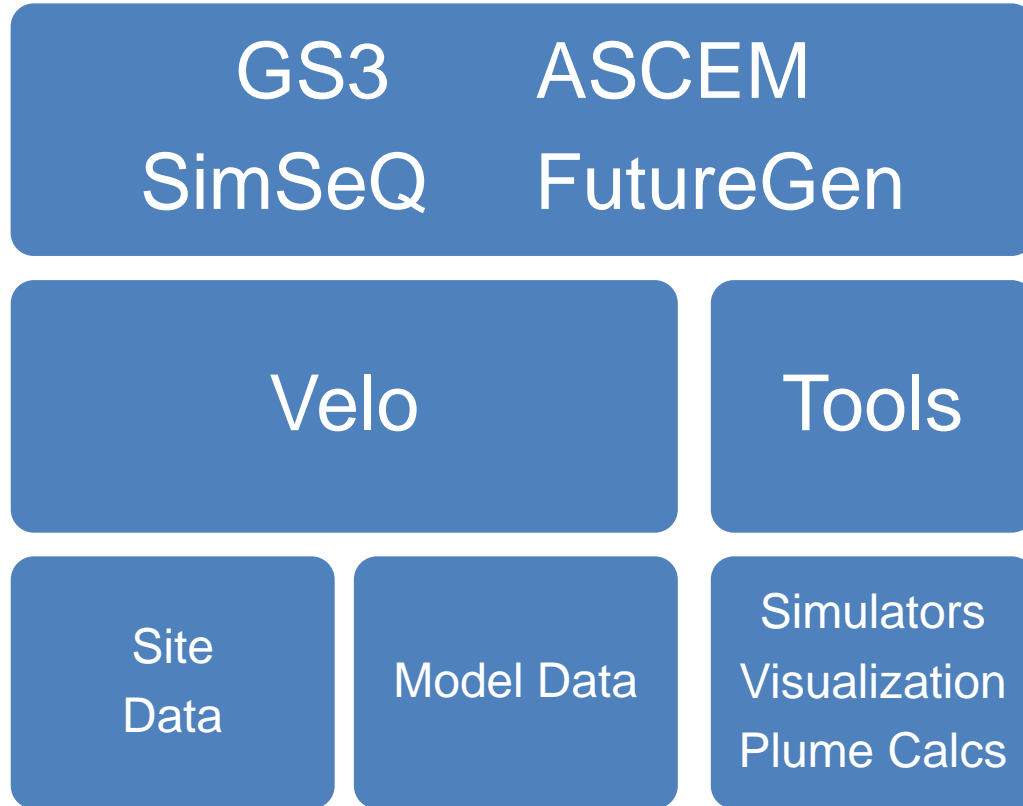
OK Cancel

# What Happened?

- ▶ Iterative development process
  - Design, build and demo, repeat
- ▶ Interest from user community was strong
  - Power of mock-ups and prototypes
- ▶ New funding obtained
- ▶ Initial sites deployed
- ▶ And along the way ...



# Flexible, Rigorous Scientific Knowledge Management



User customizable 'skins'  
Web-based  
Extensible

Raw data and metadata storage  
Versioning  
Provenance  
Tool registry  
Many deployment options

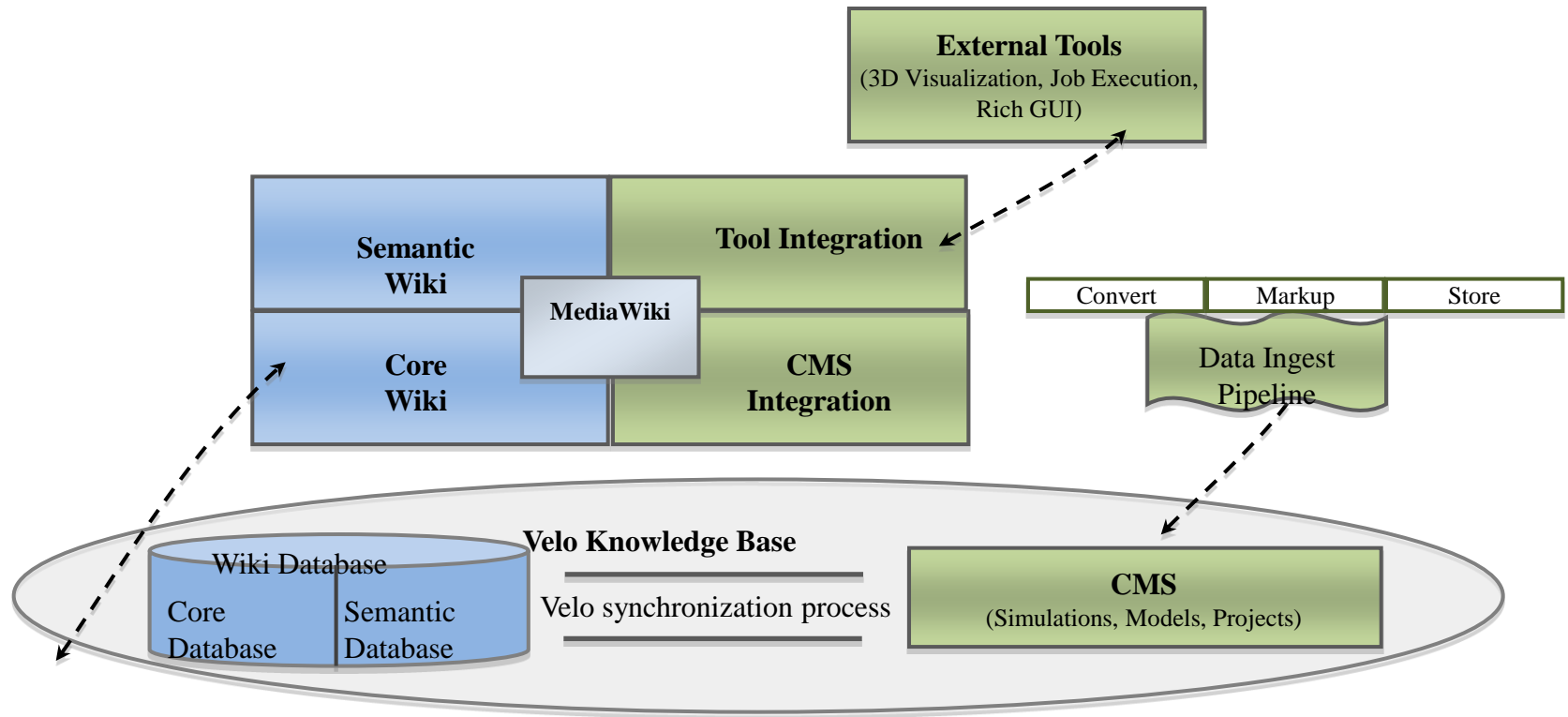
Extensible data types  
Extensible tool repository  
Programming interfaces



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Velo Architecture



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Basic Velo 'Skin'

The screenshot shows the 'Velo Demonstration' web application. At the top, it says 'Currently Logged In as: Devuser | Logout'. Below this is a navigation bar with links: Home, Browse, Tools, Misc. Tools, and Account Links. On the left side, there is a 'ScratchPad' and a 'WFS Browser' section. The main content area displays the path '/sites > UserLogin > /users/Devuser > /projects >' and a set of tools: wfs, Discussion, Annotate, Edit With Form, Edit, History, Watch, and Purge. The title of the page is 'WFS:/projects/Demo project'. Below the title, there is a table with two columns: one for location details and one for statistics. The location details column shows 'No valid location to display map' and a 'Category: Project' link. The statistics column lists various counts, all of which are zero. At the bottom, there is a 'Powered By' logo for MediaWiki and a timestamp: 'This page was last modified on January 21, 2011, at 21:55.'

/projects/Demo project	
Coordinates	No. of sites
City	No. of projects
State	No. of models
Country	No. of simulations
	No. of collections

1. Tool Navigation
2. File Browser
3. Wiki Functions
4. Content Viewer



# When requirements collide

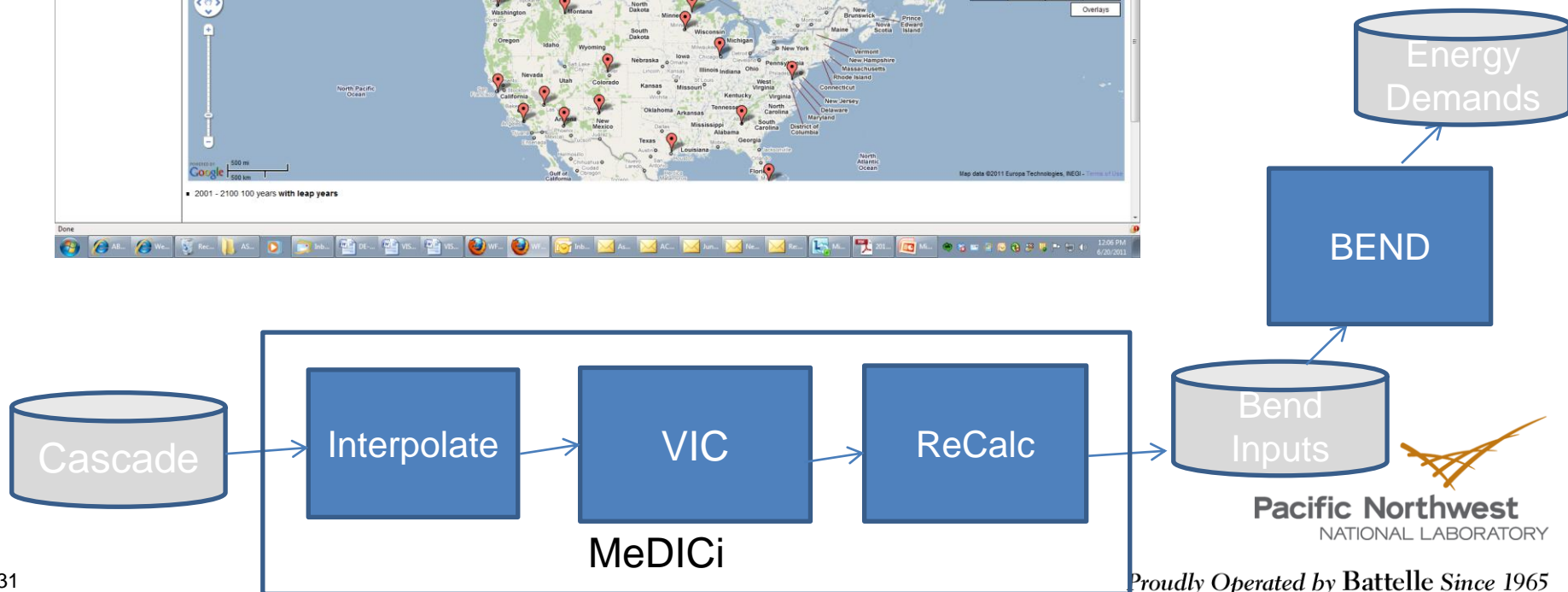
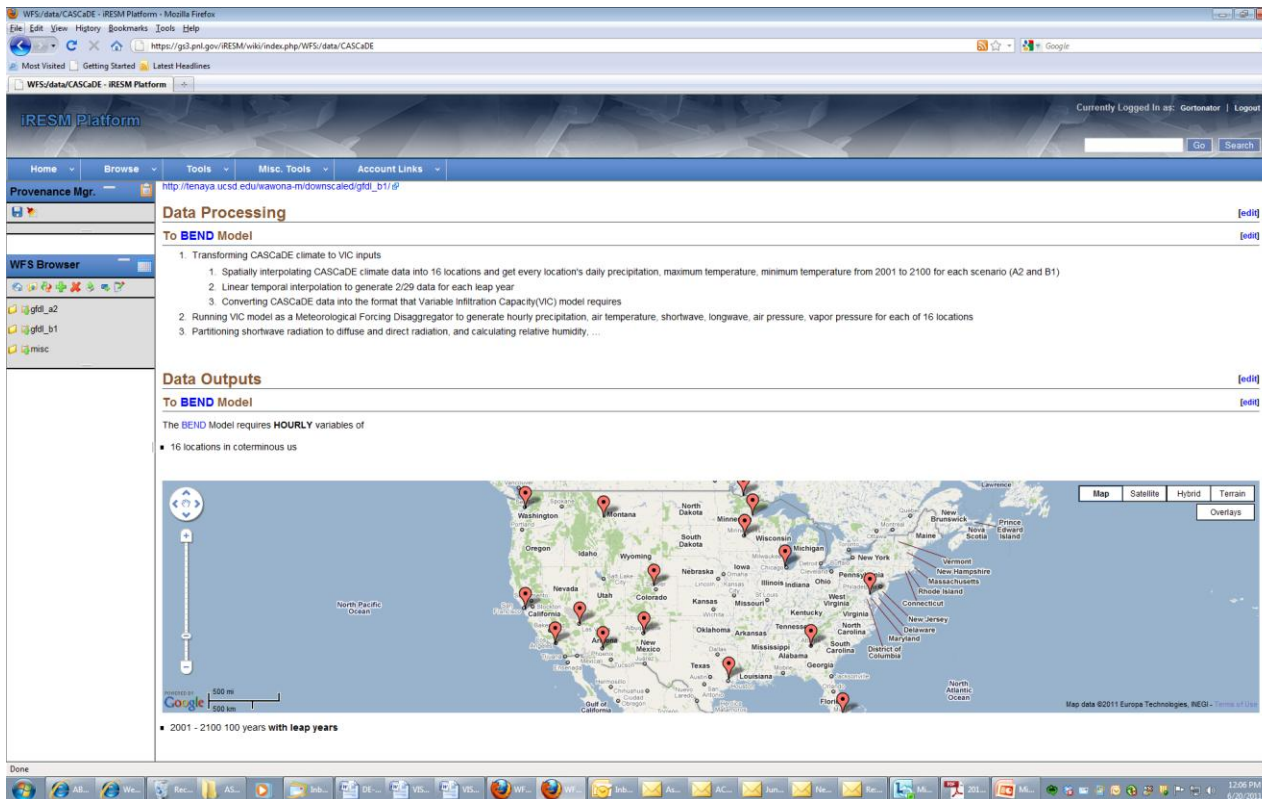
Serendipity is  
not an  
accident



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Bringing it all Together – iRESM Platform



# iRESM Feature – Reuse Matrix

Feature	Solution	Notes	Reuse
Collaboration	Velo	Core wiki features support this, includes support for discussions	100%
Sharing data	Velo	Requires integration of MW and Alfresco	100%
Metadata extraction	Velo	Requires creation of data-type specific extraction pipelines	50%
User-driven customization	Velo	Core wiki features support this	100%
Linking to model execution	Velo MeDICi	Need to construct MeDICi pipelines and link to Velo	70%
Model versioning	Velo	Supported	100%
Data set versioning	Velo	Supported	100%
Role-based Security	Velo	Supported	100%

# Some reflections

- ▶ Science is a complex domain
  - Understanding requirements
  - Diversity of software/data
  - Users who are pushing the boundaries
  - Scientists don't (in general) understand complexity of software systems
  - Architectures, integration, testing
  - Different to implementing a set of equations
- ▶ Through deliberate, creative reuse and a strong focus on architecture, we've:
  - Built generically useful technologies at low cost)
  - They work ;)



**Pacific Northwest**  
NATIONAL LABORATORY

*Proudly Operated by Battelle Since 1965*

# Questions?

